Data Mining for Climate Model Improvement

Amy Braverman

Jet Propulsion Laboratory,
California Institute of Technology
Mail Stop 126-347
4800 Oak Grove Drive
Pasadena, CA 91109-8099

email: Amy.Braverman@jpl.nasa.gov

Robert Pincus and Cris Batstone

Climate Diagnostics Center,
NOAA Earth System Research Laboratory
325 South Broadway, R/PSD1
Boulder, CO 80305

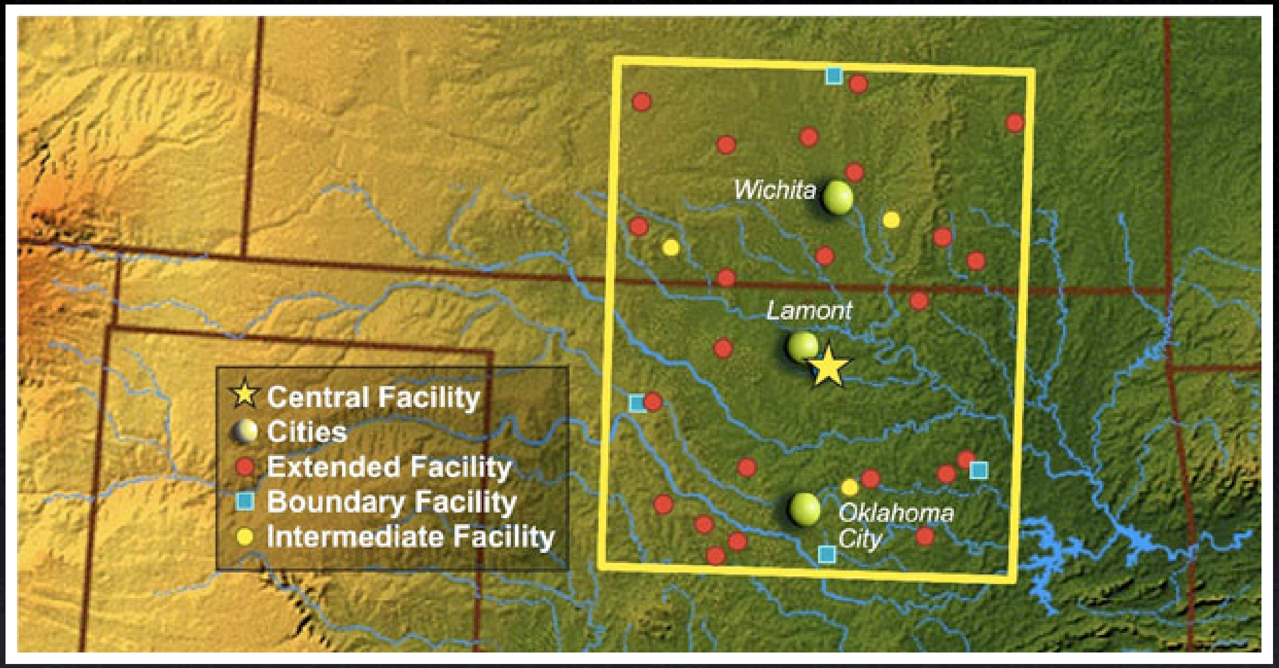
Outline

- Introduction
- Model output and observations
- Estimating multivariate distributions
- Distributional analysis
  - Visual comparisons
  - Hypothesis testing
- Conclusions

Introduction

- Model diagnosis = comparison against observations.
- Model output and observational data sets are too large to make use of.
- Instead, reduce (compress) both sources to multivariate distribution estimates; compare distributions.
- Use tools of statistics and elementary probability to characterize discrepancies.
- Work in progress!

Model Output and Observations



- Study area: Southern Great Plains (SGP) ARM (Atmospheric Radiation Measurement Program) site (north-central Oklahoma).

- Observations: vertical profiles of equivalent potential temperature (θ_e), equivalent saturation potential temperature (θ_{es}) at 35 atmospheric levels, every 30 minutes 1999-2001.

- Model output: GFDL (Geophysical Fluid Dynamics Laboratory's AM2 atmospheric model) vertical profiles of the same variables for the $2.5^\circ \times 2.5^\circ$ grid box containing the SGP site, at the same levels, every 20 minutes 1999-2001.

Model Output and Observations

$\mathbf{x}_{t_1,A}$ = 35 measurements (levels) of θ_e and 35 measurements of θ_{es} at time t_1 for ARM.

$\mathbf{x}_{t_2,G}$ = 35 measurements (levels) of θ_e and 35 measurements of θ_{es} at time t_2 for GFDL.

1:00:30 1:01:00 1:01:30

$\mathbf{x}_{t_{11},A}, \mathbf{x}_{t_{12},A}, \mathbf{x}_{t_{13},A}, \ldots$

Temporal mismatch

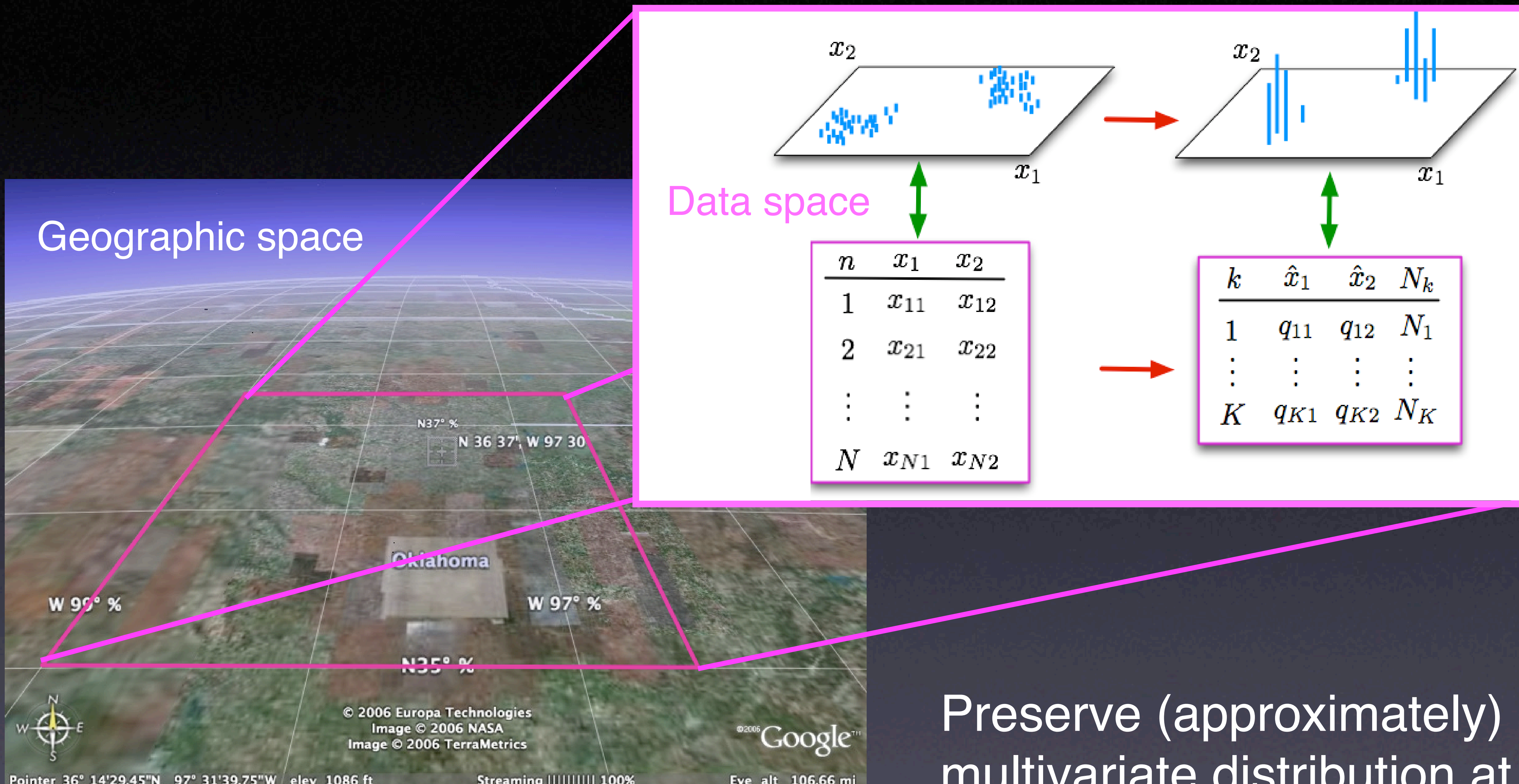
$\mathbf{x}_{t_{21},G}, \mathbf{x}_{t_{22},G}, \mathbf{x}_{t_{23},G}, \ldots$

1:00:20 1:00:40 1:01:00

How to compare?

Interpolate?
Aggregate?
Decimate?

Estimating Multivariate Distributions



Preserve (approximately) multivariate distribution at coarse spatial scale.

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Estimating Multivariate Distributions

- Entropy-constrained vector quantization (ECVQ; Chou, Lookabaugh and Gray, 1989) modified for use as a data summarization algorithm.

- ECVQ can be seen as a clustering algorithm similar to K-means. Different loss function:

$$L = \frac{1}{N}\sum_{n=1}^{N}\left[\|\mathbf{x}_n - y(\mathbf{x}_n)\|^2 + \lambda\left(-\log\frac{N_{y(\mathbf{x}_n)}}{N}\right)\right]$$

$\mathbf{x}_n$ = multivariate data point

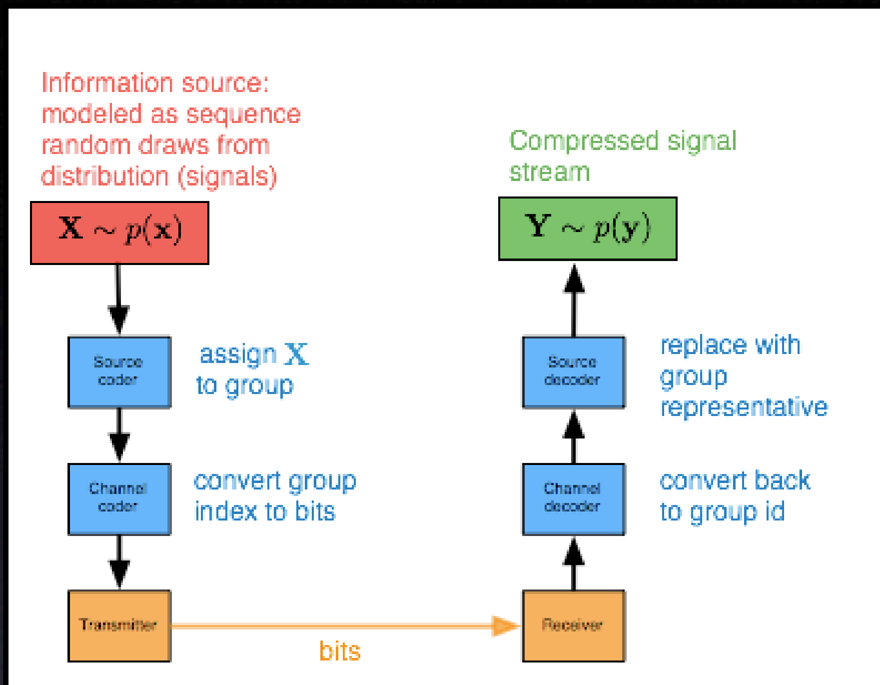
$y(\mathbf{x}_n)$ = centroid of cluster to which data point is assigned

$N_{y(\mathbf{x}_n)}$ = number of data points assigned to cluster with centroid $y(x_n)$

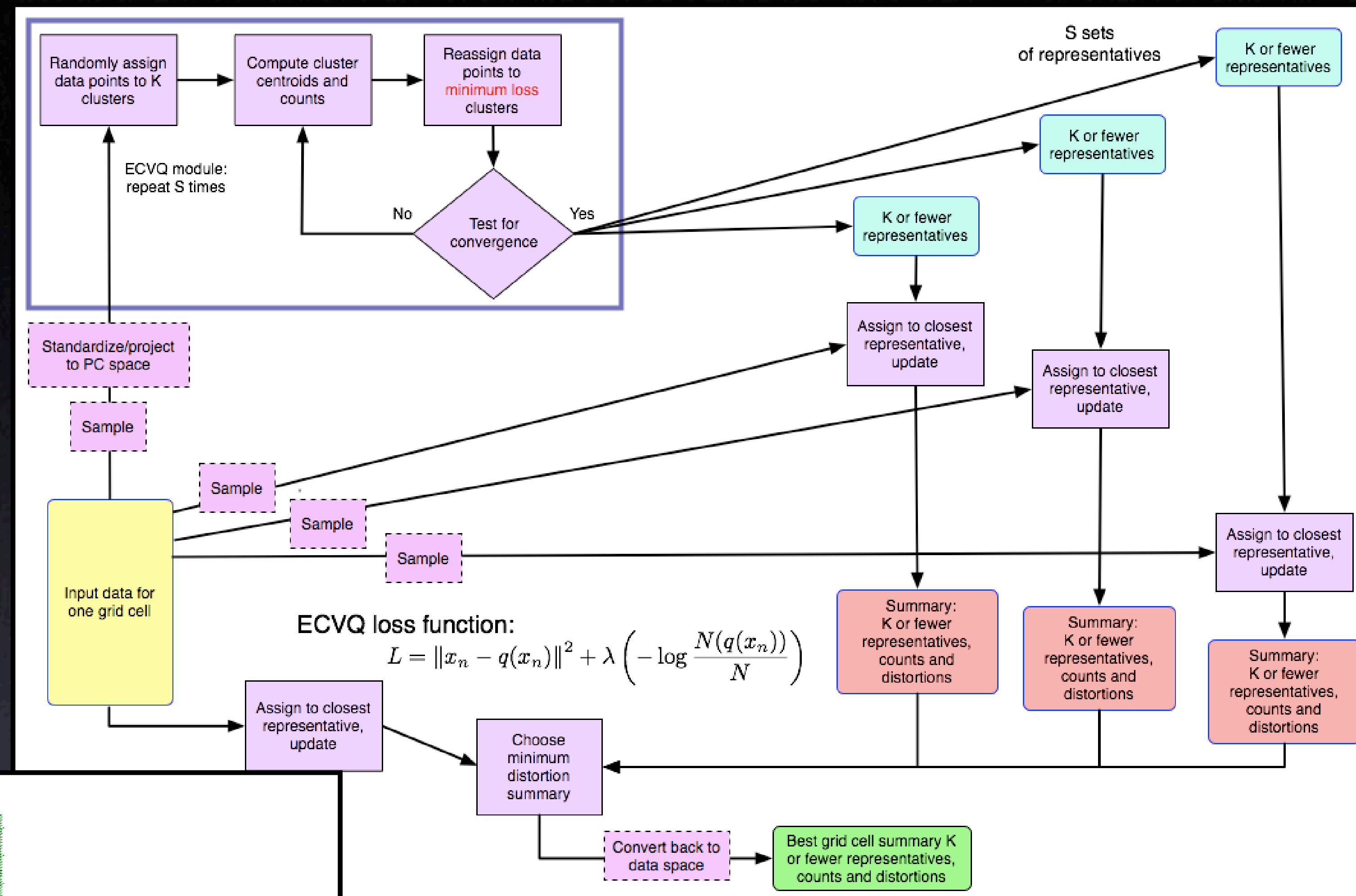
- Result: only as many clusters as necessary to describe the data, up to a maximum of K. (K-means always uses all K clusters.) Information-theoretic complexity of the data determines how many clusters.

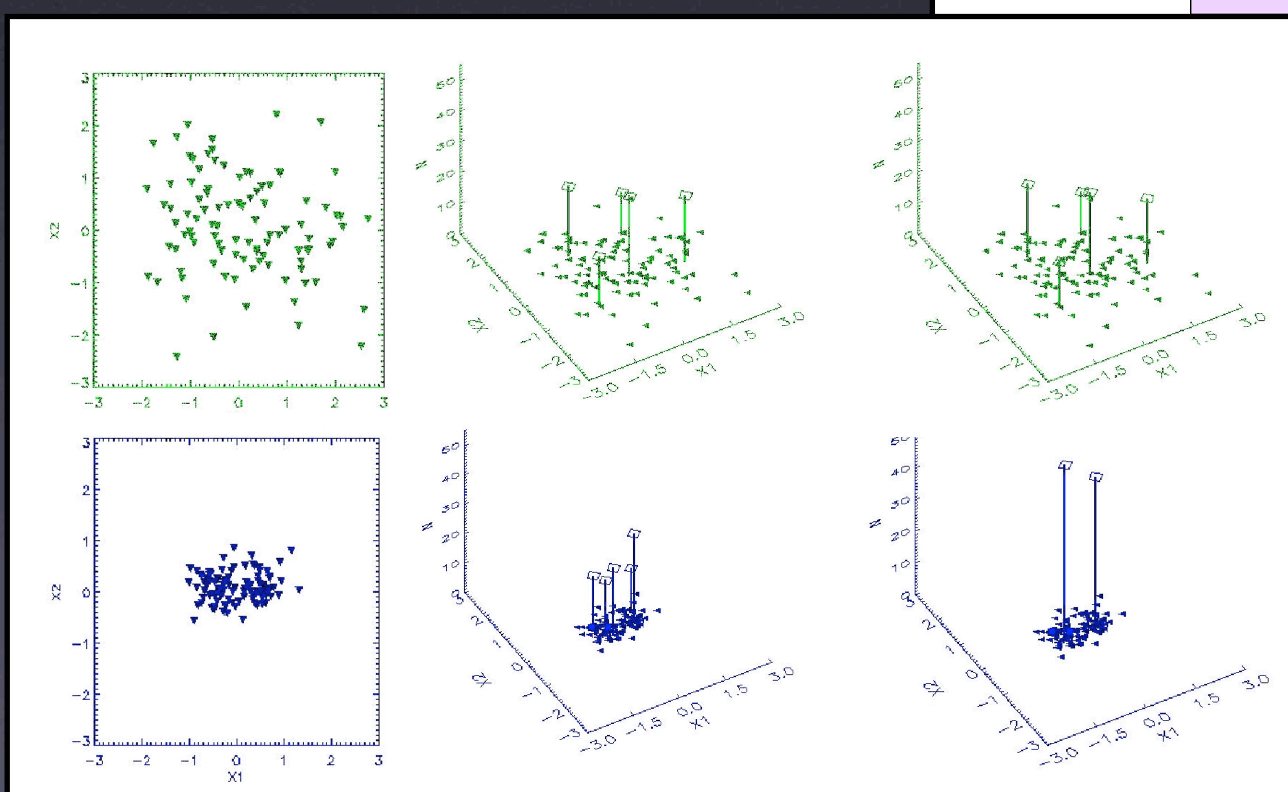
- Strategy: apply ECVQ clustering to data in grid cell(s). Produces a set of cluster centroids and weights for each grid cell.

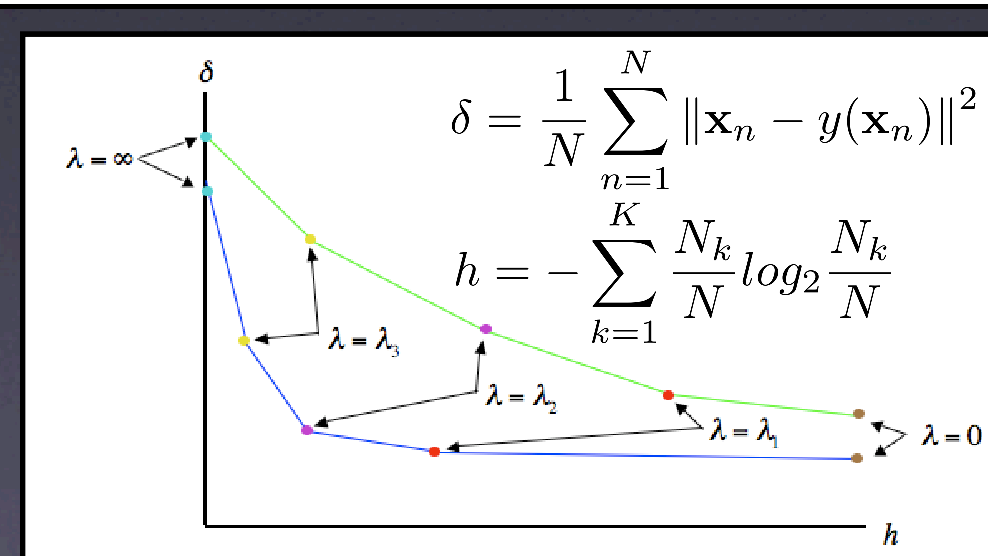
Estimating Multivariate Distributions



Signal processing paradigm







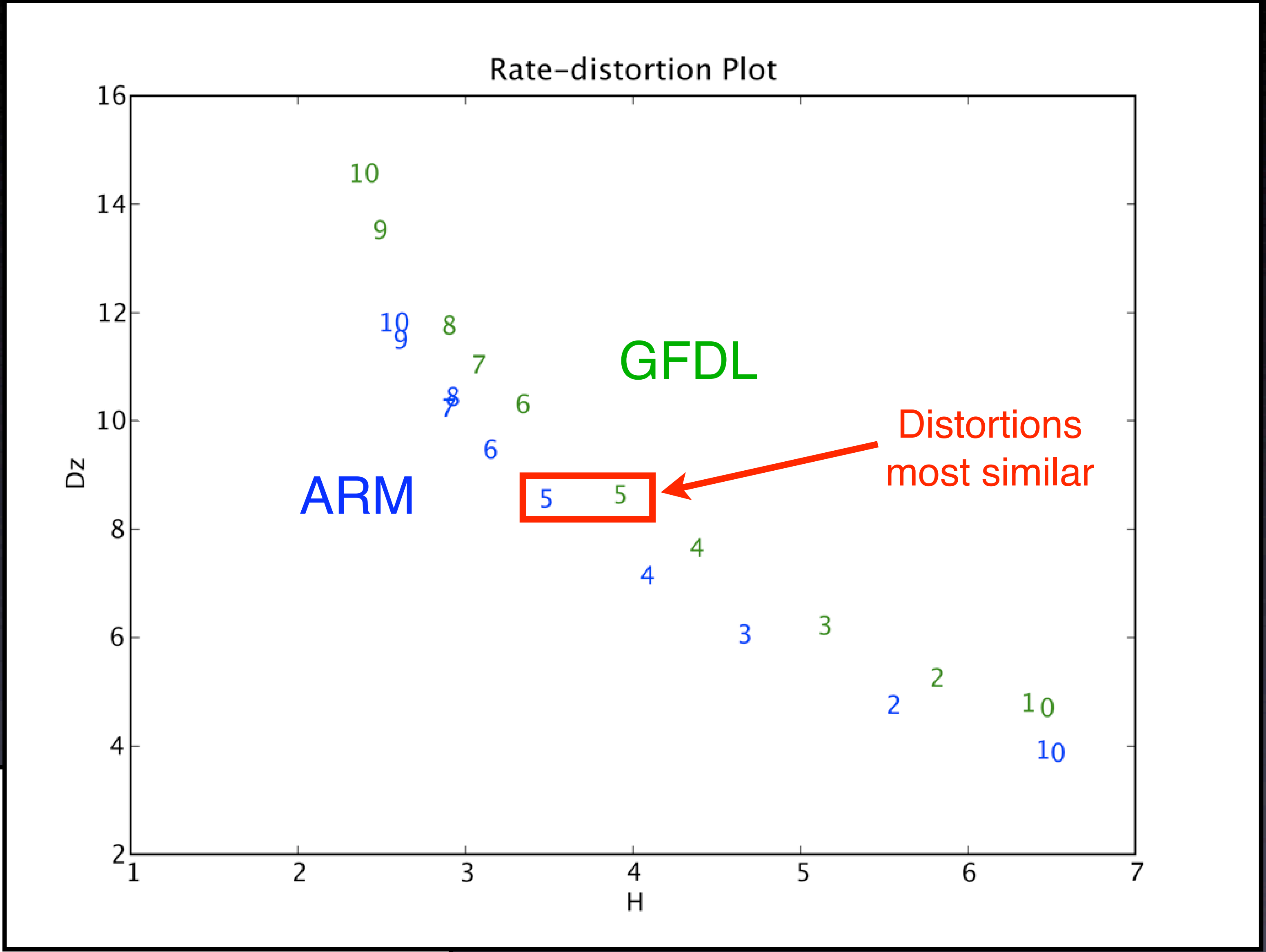
Which λ?

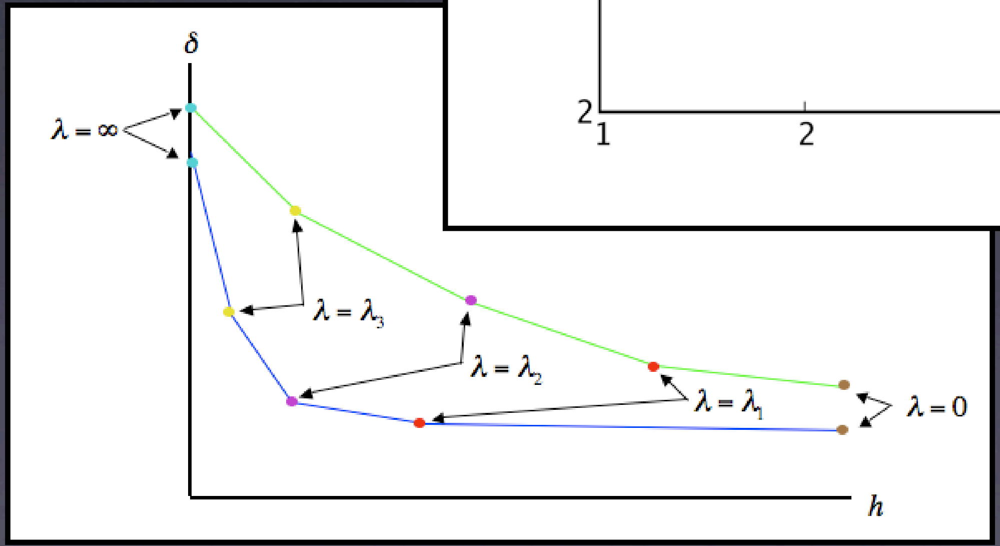
Distributional Analysis
Visual Comparisons

GFDL is more "complex":

Same accuracy requires greater entropy.

Same entropy suffers greater distortion.



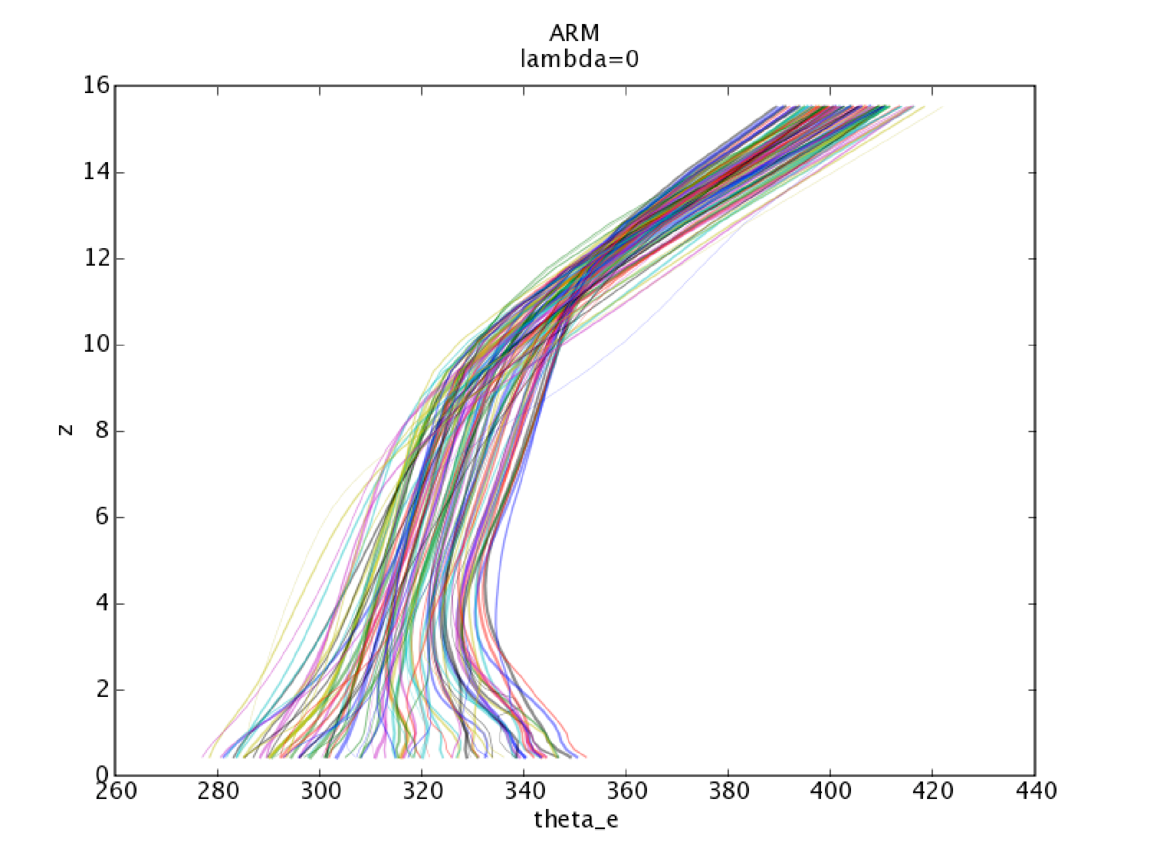


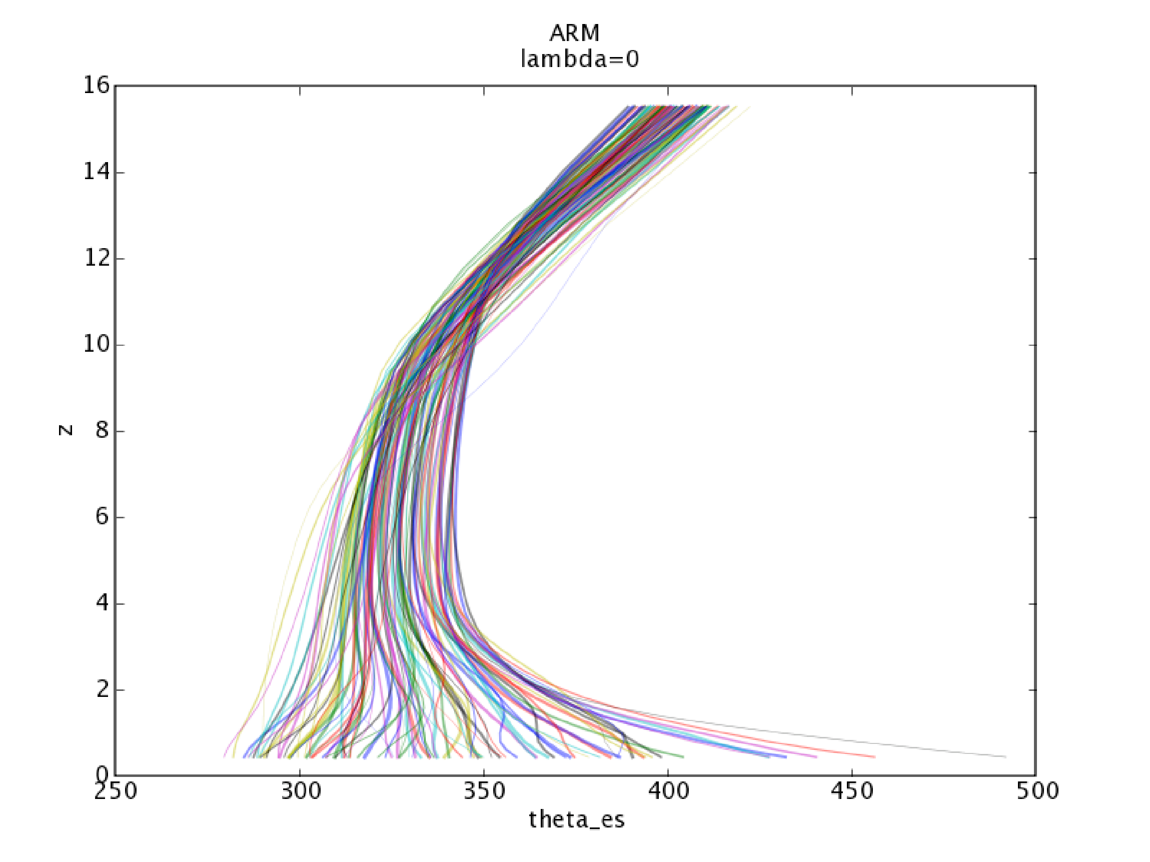
Rate-distortion plots for ARM and GFDL.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Distributional Analysis
Visual Comparisons





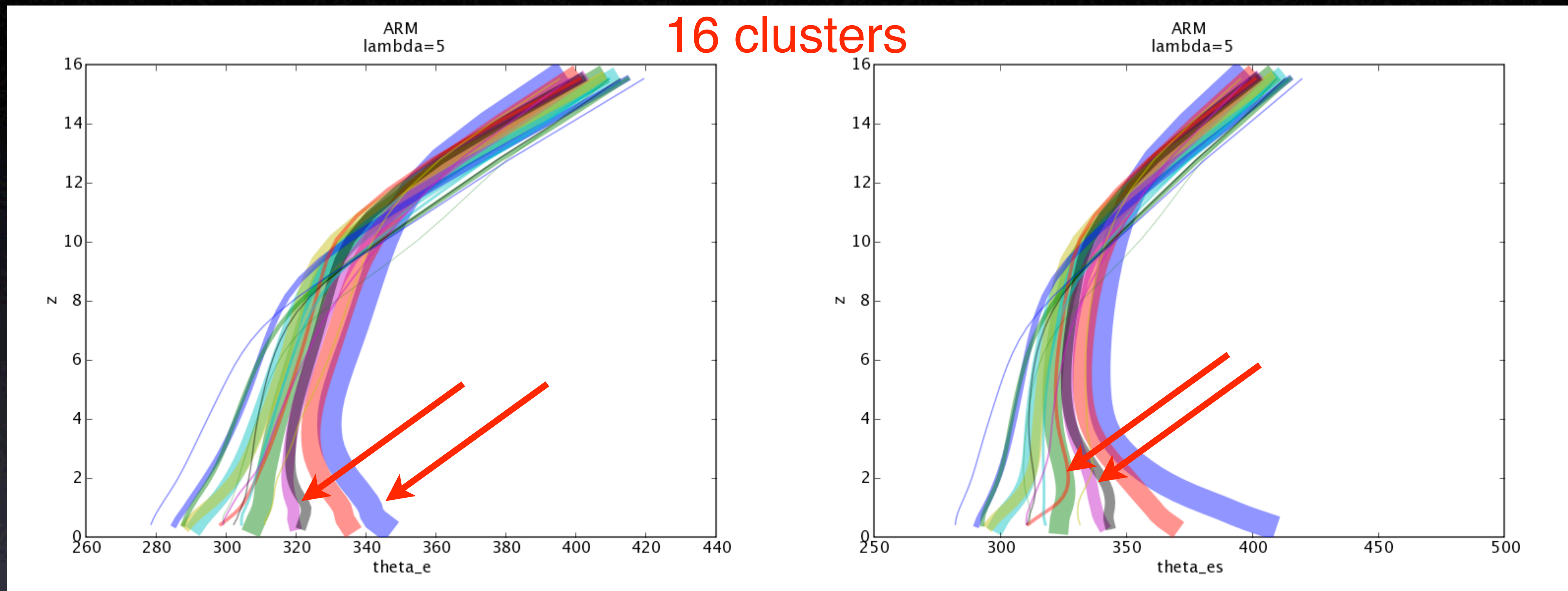
National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Distributional Analysis
Visual Comparisons

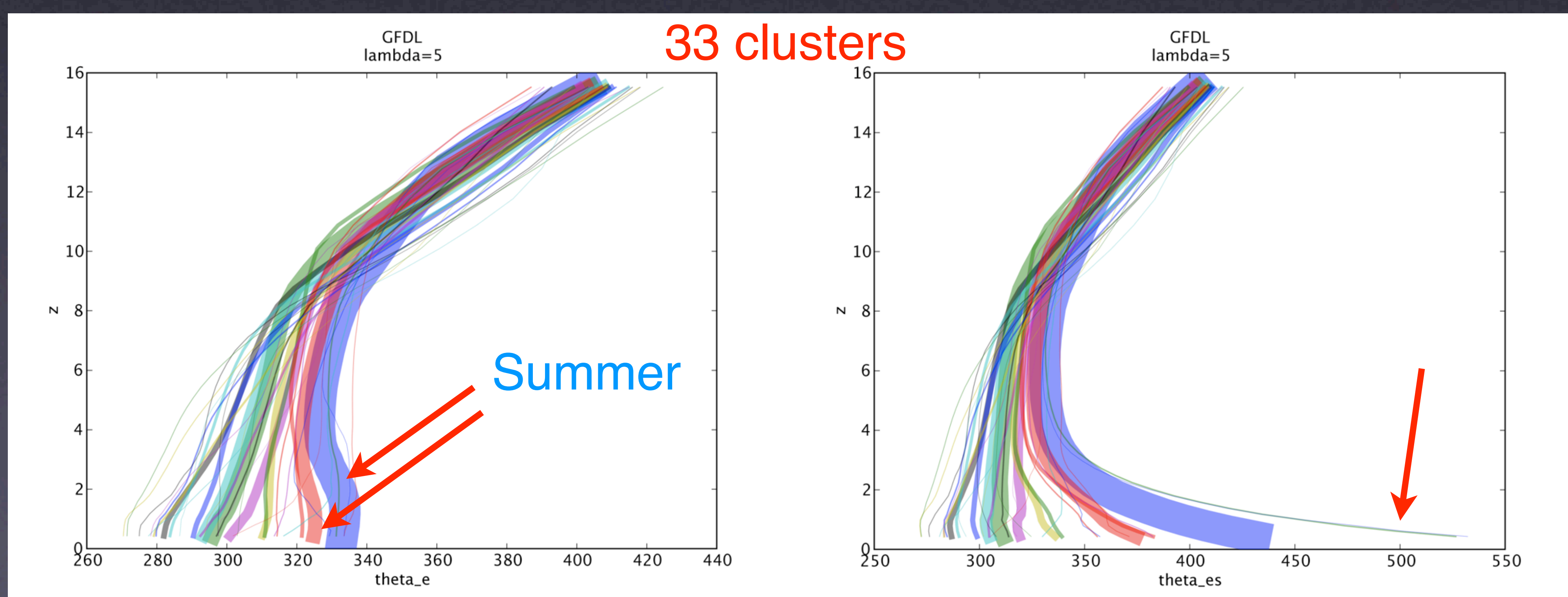
ARM θ_e

ARM θ_{es}



GFDL θ_e

GFDL θ_{es}



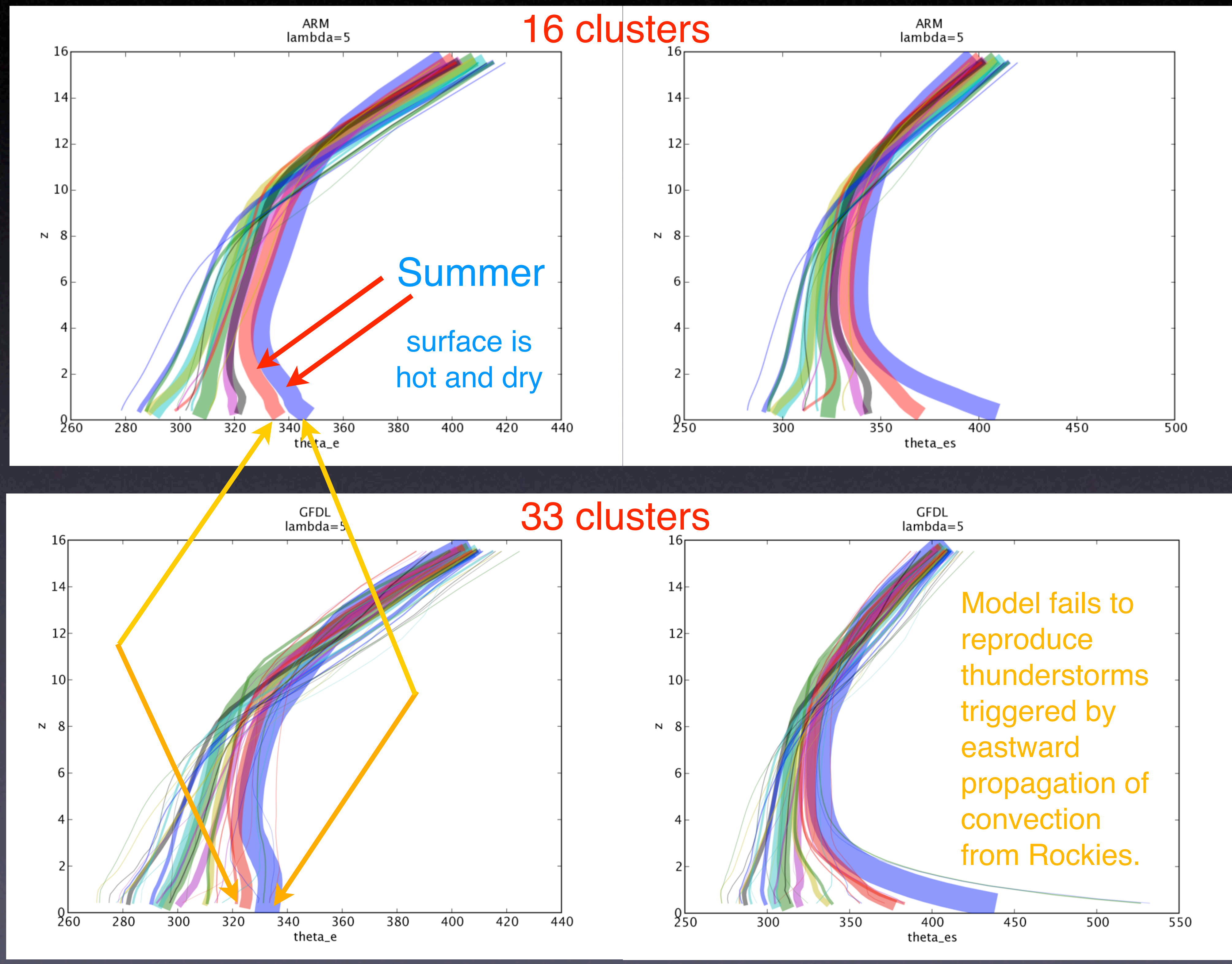
Distributional Analysis
Visual Comparisons

ARM θ_e

ARM θ_{es}

GFDL θ_e

GFDL θ_{es}



Distributional Analysis: Hypothesis Testing

- Are the distributions of ARM and GFDL the “same”?
- Test the hypothesis that the GFDL distribution (P_2) could have been obtained by sampling from a population that looks like the ARM distribution (P_1).
  - Formulate a test statistic that measures the extent to which two distributions differ ($\Delta(P_1, P_2)$).
  - Do the following 100 times:
    - draw N data points randomly from the ARM distribution;
    - cluster them to produce $P_1^*, P_2^*, \ldots, P_{100}^*$;
    - calculate $\Delta_b^* = \Delta(P_1, P_b^*)$, the similarity between P_1 and P_b^*;
    - make a histogram of the Δ_b^* ‘s, $b = 1, 2, \ldots, 100$;
  - If less than 5% of the histogram is greater than the actual $\Delta(P_1, P_2)$, then reject the hypothesis (at the 5% significance level).

Distributional Analysis: Hypothesis Testing

- A distance between distributions:

$$\pi_1 = \{(y_{1k_1}, \pi_{1k_1})\}_{k_1=1}^{K_1} \qquad \pi_2 = \{(y_{2k_2}, \pi_{2k_2})\}_{k_2=1}^{K_2}$$

$$\Delta(\pi_1, \pi_2) = \min_{p_{12}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \|y_{1k_1} - y_{2k_2}\|^2 p_{12}(y_{1k_1}, y_{2k_2})$$

column constraints

$$\pi_{11} = p_{11} + p_{21} + p_{31}$$
$$\pi_{12} = p_{12} + p_{22} + p_{32}$$
$$\pi_{13} = p_{13} + p_{23} + p_{33}$$
$$\pi_{14} = p_{14} + p_{24} + p_{34}$$

π 's are fixed; fill in p 's such that:

(1) constraints are satisfied
(2) Δ is minimized

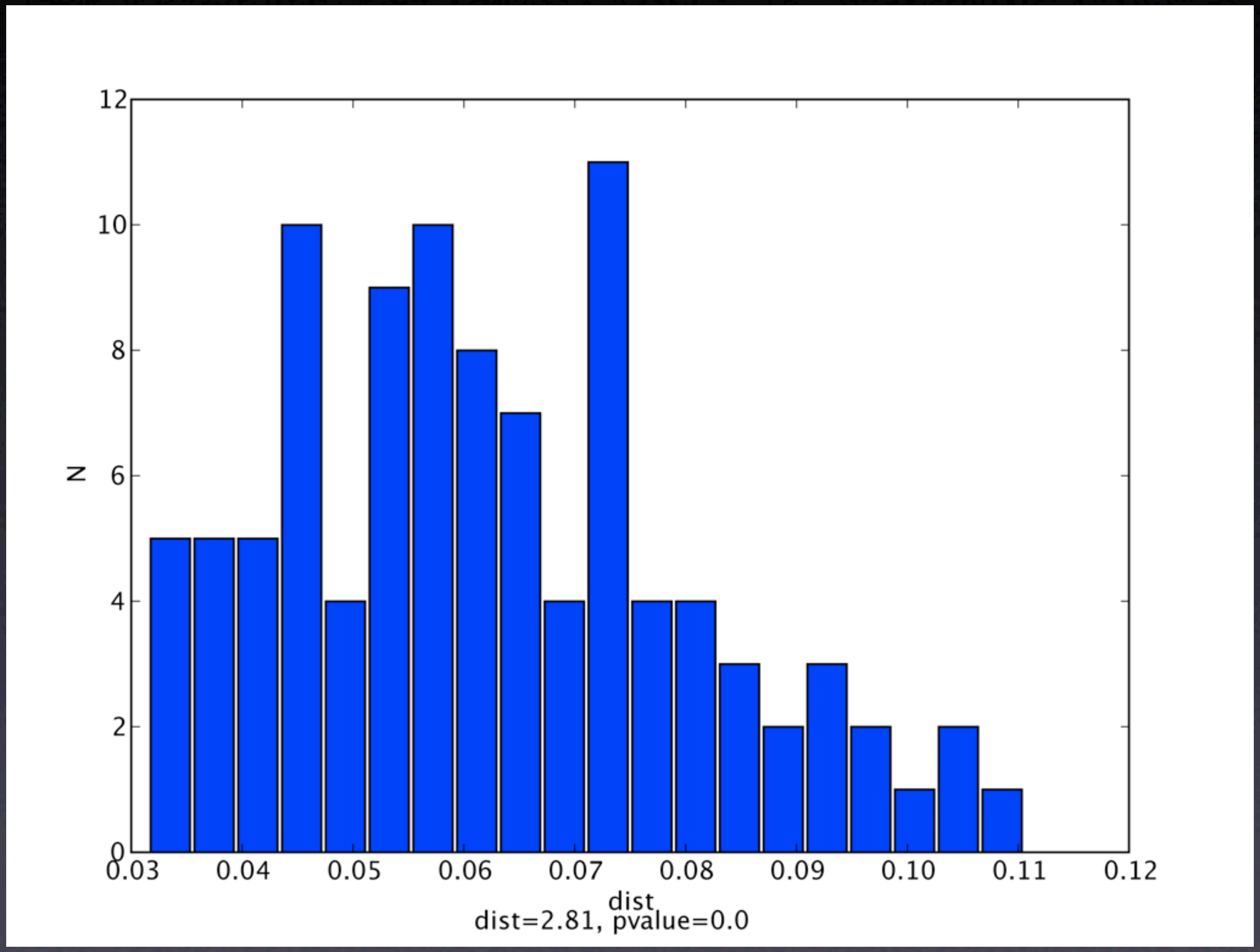
$$\pi_{21} = p_{11} + p_{12} + p_{13} + p_{14}$$
$$\pi_{22} = p_{21} + p_{22} + p_{23} + p_{24}$$
$$\pi_{23} = p_{31} + p_{32} + p_{33} + p_{34}$$

row constraints

		y_{11} π_{11}	y_{12} π_{12}	y_{13} π_{13}	y_{14} π_{14}
y_{21}	π_{21}	p_{11}	p_{12}	p_{13}	p_{14}
y_{22}	π_{22}	p_{21}	p_{22}	p_{23}	p_{24}
y_{23}	π_{23}	p_{31}	p_{32}	p_{33}	p_{34}

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Distributional Analysis: Hypothesis Testing



Histogram of Δ_b^*

Actual $\Delta(P_1, P_2) = 2.81$

Reject the hypothesis; ARM and GFDL distributions are not the same to within sampling variability.

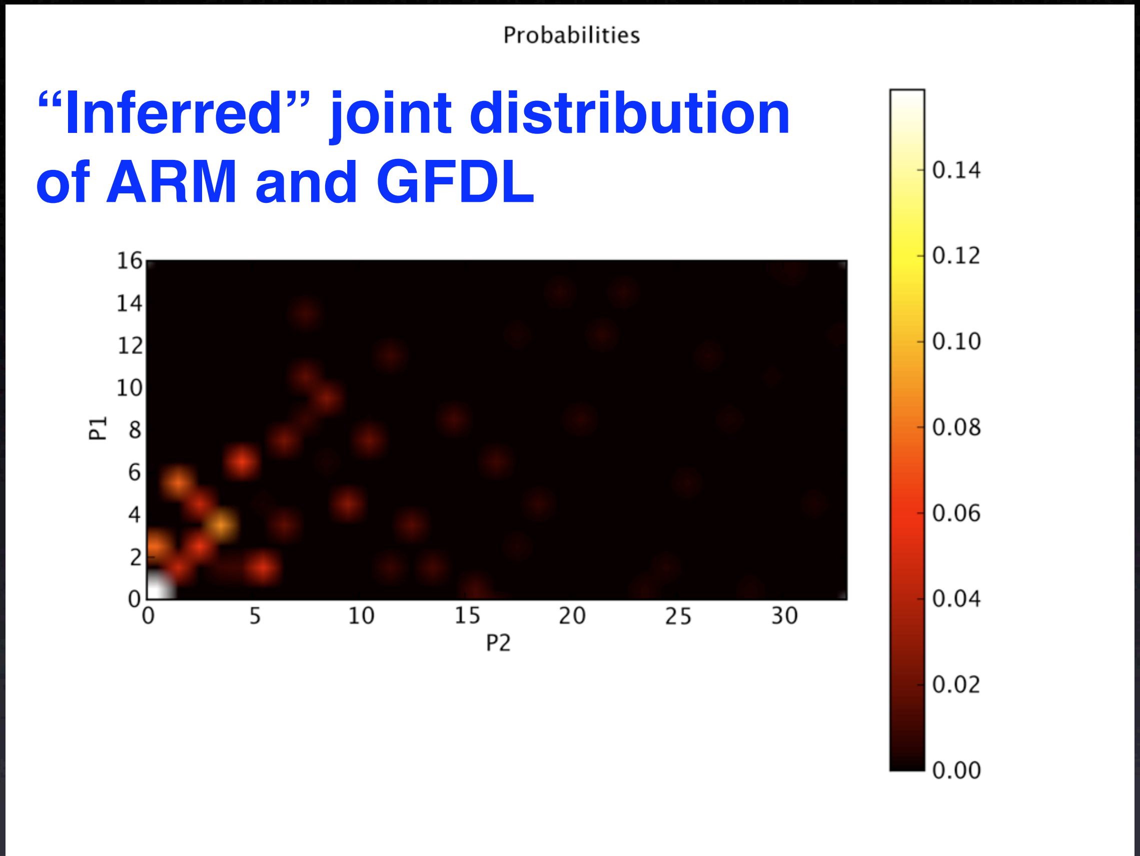
Why?

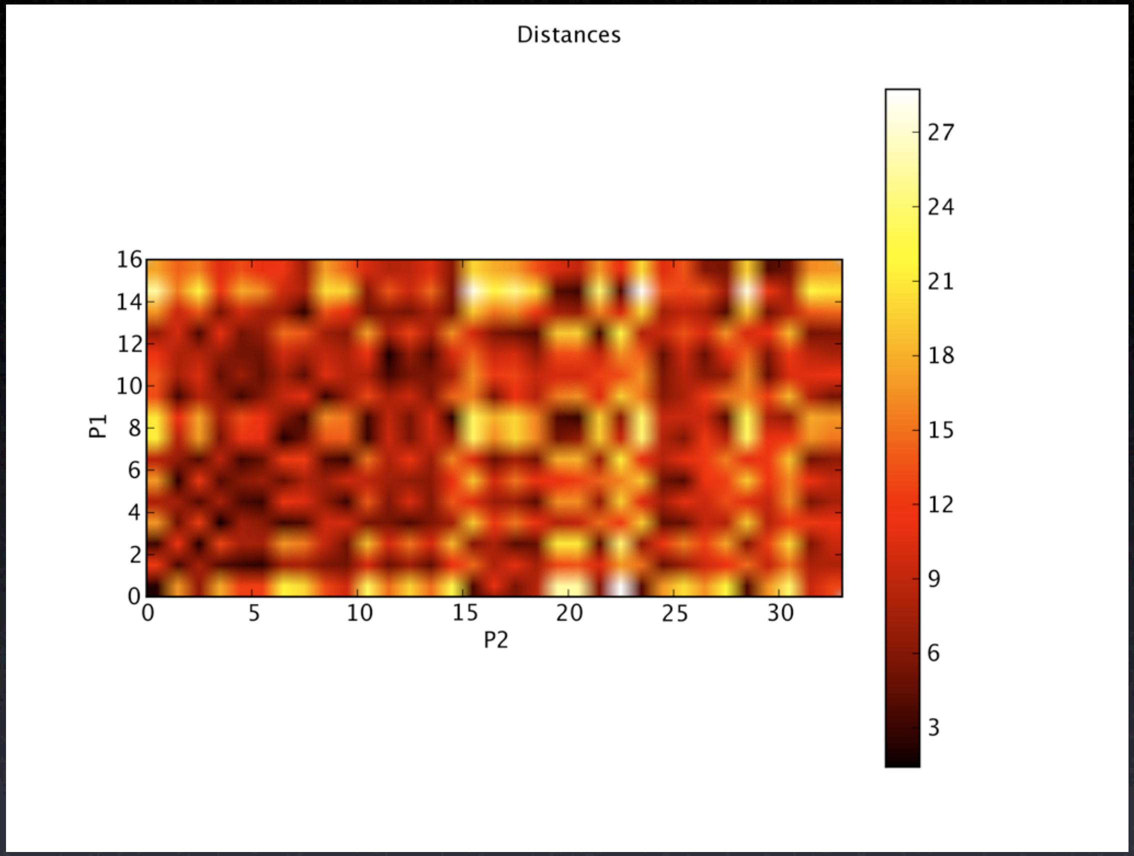
Which parts of the distribution lead to rejection?

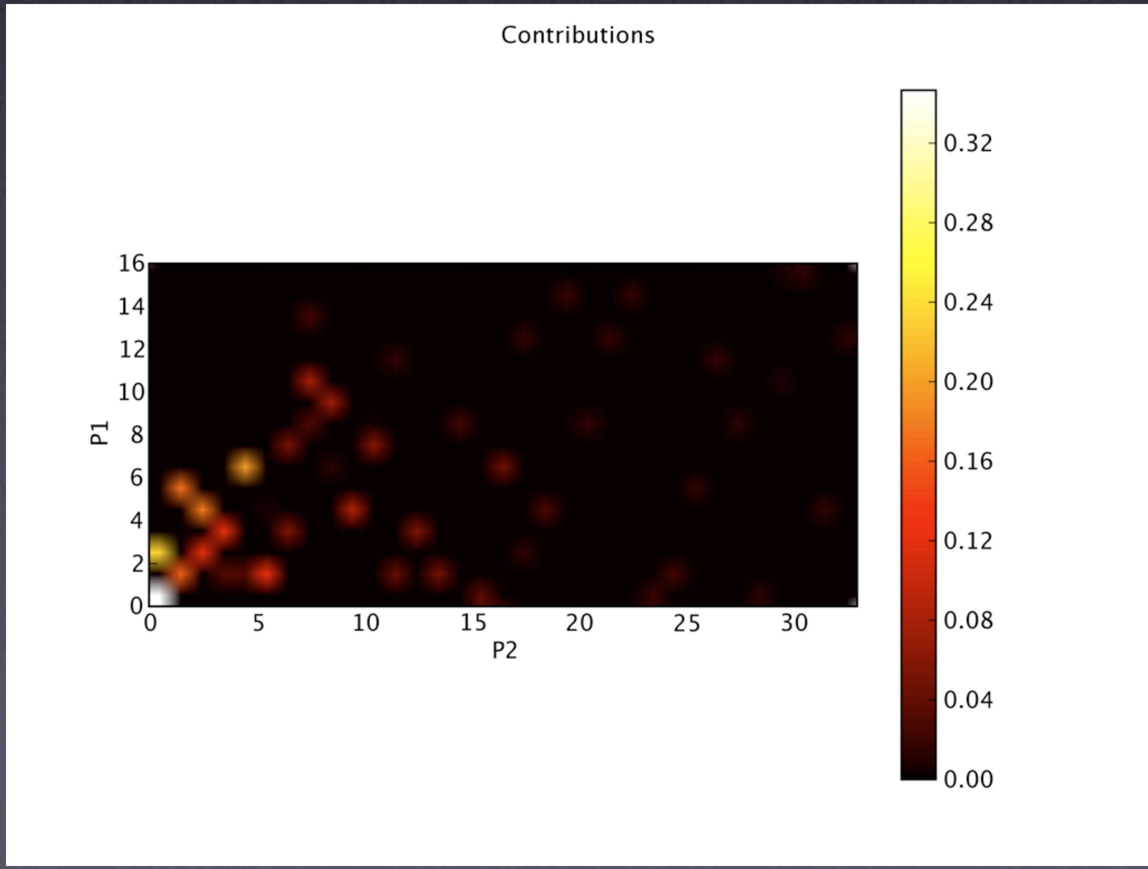
What physical processes do they correspond to?

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Distributional Analysis: Hypothesis Testing







Largest contributions to $\Delta(P_1, P_2)$ do not correspond to largest distances.

Shows how difficult the problem is!

Distributional Analysis: Hypothesis Testing

An Alternate Approach

- Each cluster represents a distribution of values with mean vector = representative and dispersion = distortion.

- Markov's Inequality bounds the probability of an observation being more distant from the mean than a given amount:

$$P(X > a) \leq \frac{EX}{a}, \quad X = \|\mathbf{X} - y(\mathbf{X})\|^2 \qquad \text{implies}$$

$$P(\|\mathbf{X} - y(\mathbf{X})\|^2 > 20\delta) \leq 0.05$$

Test a set of hypotheses: GFDL cluster j's representative could have been drawn at random from ARM cluster i's distribution...

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Distributional Analysis: Hypothesis Testing

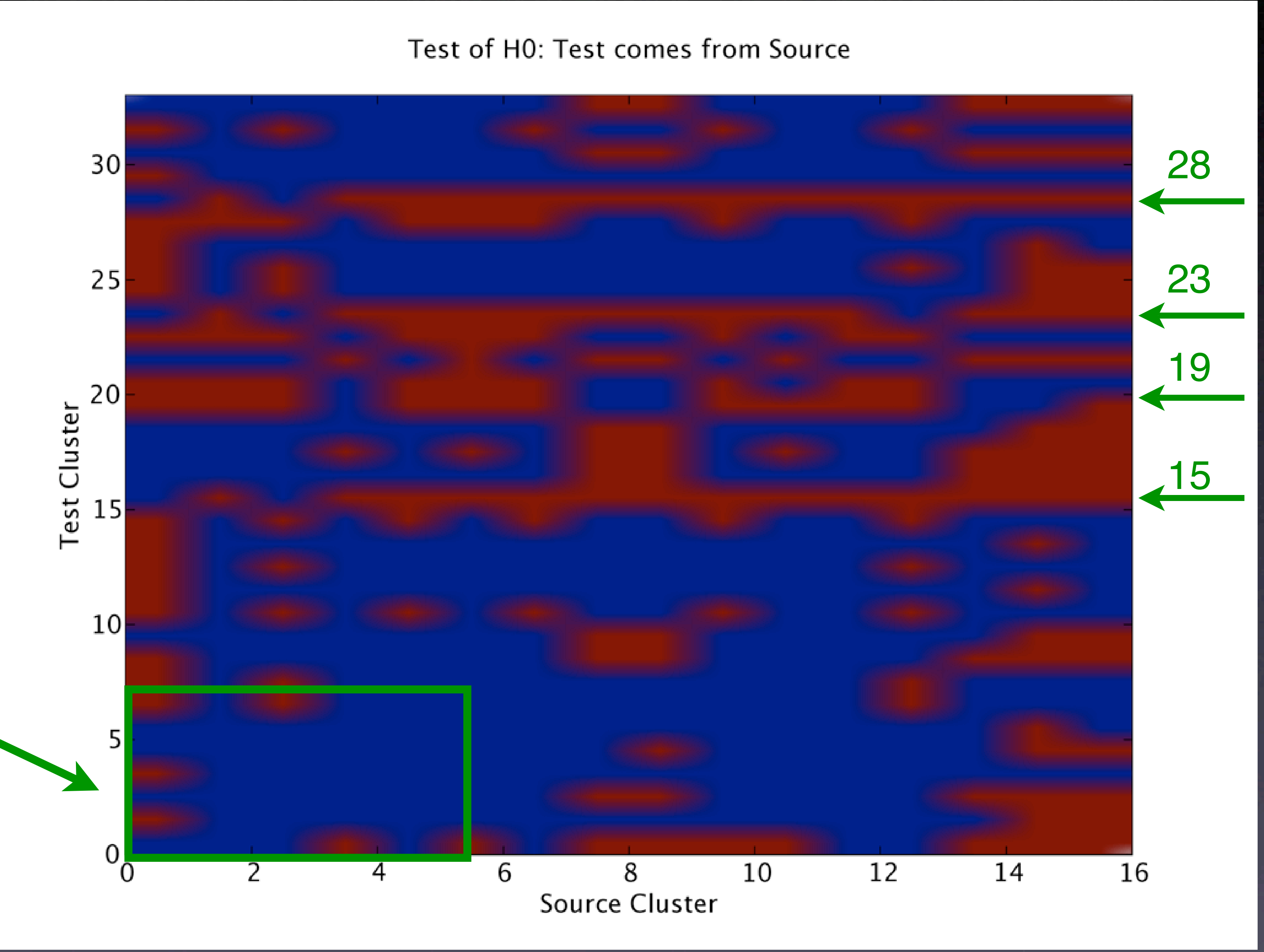
An Alternate Approach

Red=reject

Blue=do not reject

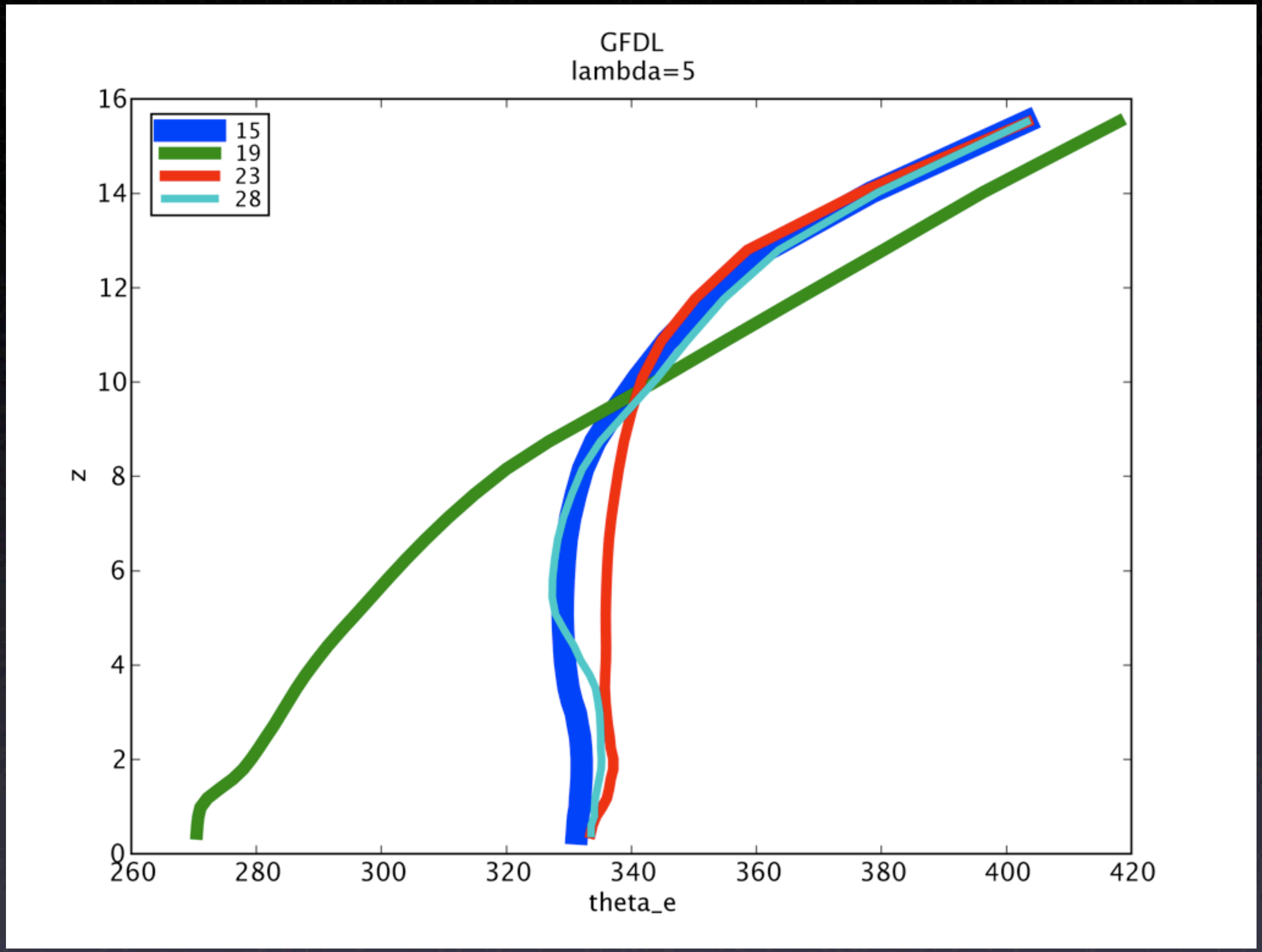
GFDL

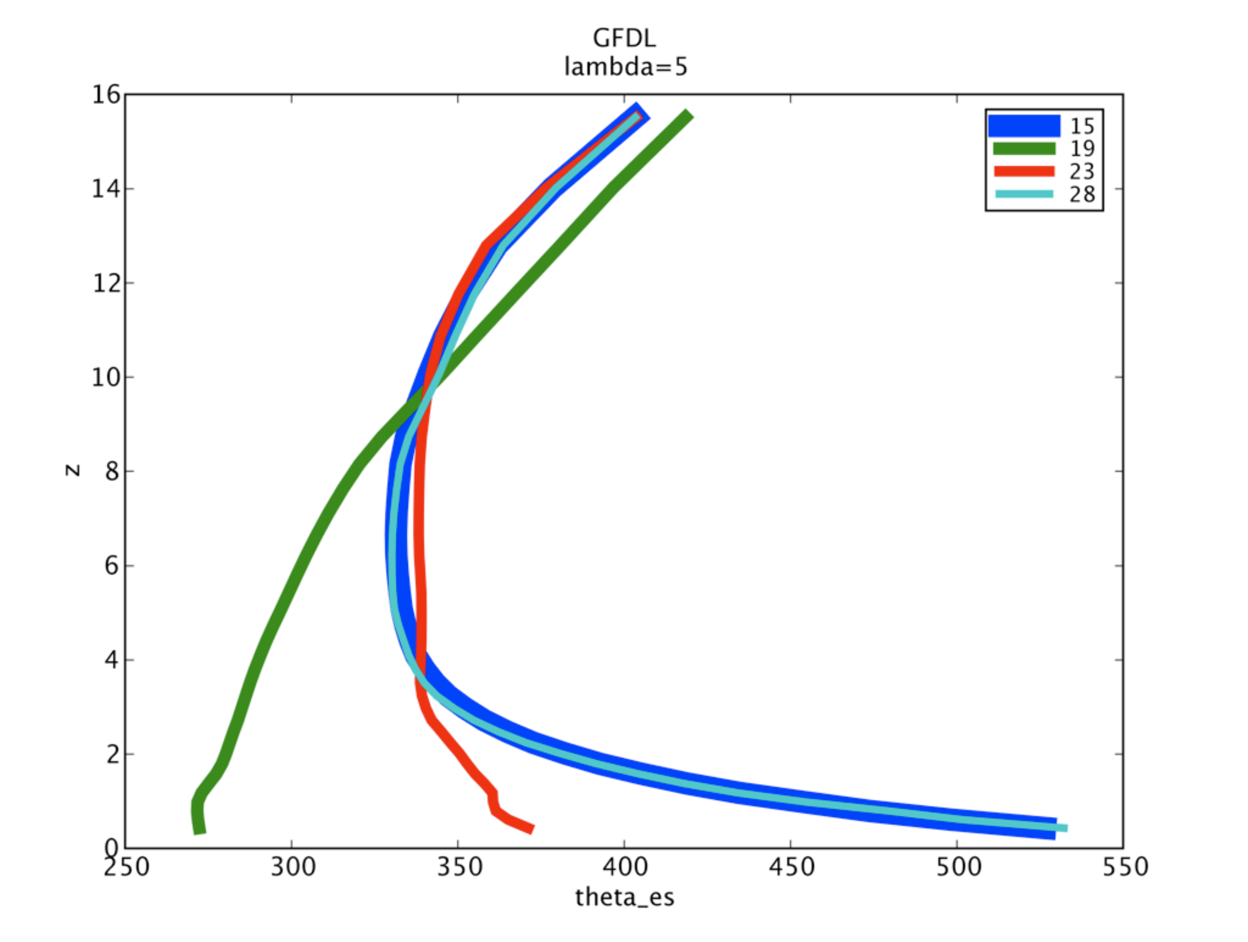
How consistent with the first approach?



ARM

Distributional Analysis: Hypothesis Testing





- GFDL clusters 15 and 28 below 2 km are not physical- too hot and too dry. Precipitation not handled properly.
- GFDL cluster 19: cloudy and unrealistically stable atmosphere.
- GFDL cluster 23?

National Aeronautics and Space Administration
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Conclusions

- Problem is to discover why model output and comparable data do not agree.
- Estimate discrete multivariate data distributions and compare them to isolate sources of discrepancy.
- Visual inspection is useful, but we need an “autonomous” method suitable for large data sets.
- Two approaches to hypothesis testing using discrete distributions- mixed results, but we are not finished.
- Thanks to ESTO and the AIRS and MISR projects their for support!